Lots Of Copies Keep Stuff Safe: Peer-to-Peer Digital Preservation

David S. H. Rosenthal

Stanford University Libraries
http://lockss.stanford.edu

© 2004
No Redistribution Without Permission

Status

- Libraries can preserve copyright e-journals
  - Cooperate to audit, detect and repair damage
- Five years of testing ended April 2004
- In production use at ~80 libraries worldwide
- Publishers of 2000+ titles endorse system
- Light archive - content always accessible
  - No trigger events, no phase changes
- Transparent on-access format migration
- Conforms to OAIS, can ingest via OAI-PMH

Archive or Library?

…let us save what remains: not by vaults and locks which fence them from the public eye and use in consigning them to the waste of time, but by such a multiplication of copies, as shall place them beyond the reach of accident.

Thomas Jefferson, 1791

LOCKSS Overview

- Library runs peer = persistent Web cache
  - Crawls web to collect content, never flushes it
  - Reader's browser proxies via cache
  - Sees publisher copy if it can, else cached copy
- Publisher adds page granting permission to
  - Collect, preserve, supply to local readers
  - Supply repairs to other libraries
- Library republishes only to its community
  - Just like paper, less threatening for publishers
- Peer audit detects & repairs damage

LOCKSS Differences

- Content is copyright & publisher decides format
  - We can't impose formats or metadata on publishers
  - We have to do what we can with what we can get
- We have to be very, very cheap for a library to use
  - For us, User Interface is a problem not a solution
- Our customer is an ordinary Web surfer
  - Not a skilled professional archivist
- Digital Preservation for the Rest of Us
  - *Make It Simple*, Andreas Kluth, *The Economist*, 10/30/04

Guard Against Failures

- Economic: bits need to be fed money
  - Risk: one budget, one cut, total failure
- Technical: hardware/software unreliable
  - And so are system operators
  - Audit essential to detect failures
- Confidence: believe archive will work?
  - Keys: open source, audit & light archive
- Attacks: system *will* be attacked
  - Firewall is illusory many attacks by insiders
- Failure must be *infrequent* and *slow*

Copyright

- Need permission to preserve copyright content
  - Even for open access content
- Must *negotiate* system design with publisher
  - Need win-win outcome
  - Priority: preserve publisher business model
    - Obvious archive design unacceptable threat
  - Trigger event another name for litigation
- Archiving own content much easier
  - But risks 1984-like history rewrite

Affordability

- Centralizing the money risks sudden collapse
  - Independent cooperating budgets more resilient
- No-one has budget to preserve everything
  - Cheaper systems can preserve more stuff
- Cheaper publisher negotiation
  - Simple blanket license, one-time negotiation
- Cheaper staff costs
  - <15 min/month, no backups, automatic audit
- Cheaper hardware
  - Reliability from replication

Auditability

- No audit, no confidence in archive operation
  - Can't just assume everything is OK
  - Can't depend on readers to report failures
- Can't recover from failures you don't detect
  - Did your web crawl get everything?
  - When do you need to restore from backup?
- Audit processes key to archive design
  - Manual audit cost can outweigh everything else
  - Mutual audit protocols support diversity

Replication & Diversity

- Replication essential to survival
- Identical replicas = instant epidemic failure
  - E.g. Slammer
- Need 3 *different* replica implementations
  - At *each* level: hardware, O/S, software
- Replicas must audit each other via a protocol
  - LOCKSS protocol is a basis for this audit
  - Attack/failure resistance won research awards

Lessons from Production

- New tool finds new uses
  - Humanities
  - Government Documents
- If humans do it, it doesn't scale
  - LOCKSS growth limited by:
    - Selection of content to preserve
    - Getting permission from publisher
  - System must be *automatic* not just automated
- If humans do it, they do it wrong
  - System must validate all human inputs

Credits, Questions?

- Funders:
  - Mellon Foundation, NSF, LOCKSS community
- Vicky Reich manages the LOCKSS program
- Engineering:
  - Tom Robertson, Tom Lipkis, Claire Griffin, Seth Morabito
- Research
  - Petros Maniatis (Intel),TJ Giuli (Stanford),
  - Mema Roussopoulos (Harvard), Mary Baker (HP)
- Download source from SourceForge